

PAGE RANK

EL GRAN EXITO DE GOOGLE FUE LOGRAR BANQUEAR LOS RESULTADOS DE LAS BUSQUEDAS DE FORMA INDEPENDIENTE DEL CONTENIDO DE LAS MISMAS MEDIANTE EL ALGORITMO DE PAGE BANK.

PAGE BANK SE BASA EN LA ESTRUCTURA DE LINKS QUE EXISTE ENTRE LOS DOCUMENTOS DE LA WEB. EL CONCEPTO BASICO ES QUE CADA PAGINA TIENE UNA CIERTA 'IMPORTANCIA' QUE ES INTRINSECA Y DEPENDE DE LOS LINKS QUE LLEVEN A DICHA PAGINA. CUANTO MAS LINKS NOS LLEVEN A UNA CIERTA PAGINA, MAS IMPORTANTE SEBA LA MISMA.

RANDOM SURFERS

ESTE ES EL MODELO MATEMATICO DETRAS DE PAGE BANK.

UN RANDOM SURFER ES UN NAVEGANTE ALEATORIO EN LA WEB. ESTE NAVEGANTE COMIENZA EN CUALQUIER PÁGINA AL AZAR DE TODAS LAS DISPONIBLES DESDE ESA PAGINA EL NAVEGANTE ELIGE UN LINK AL AZAR DE LA PAG EN LA QUE SE ENCUENTRA Y NAVEGA, REPITIENDO ESTE PROCESO INDEFINIDAMENTE EN VARIAS ITERACIONES

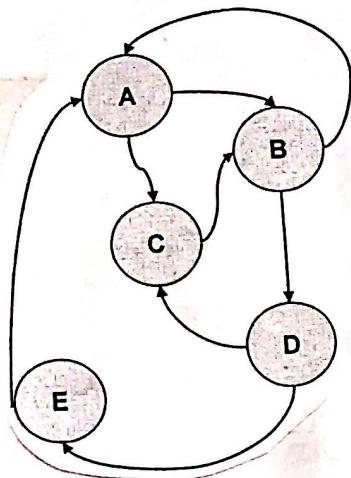


EL CONCEPTO DETRAS CONSISTE EN DARLE A CADA PAGINA UN PESO QUE ES IGUAL A LA PROBABILIDAD DE QUE NUESTRO RANDOM SURFER TERMINE SU RECORRIDO EN ESA PAG LUEGO DE N SALTOS.

TENEMOS TANTOS NODOS COMO PAGINAS Y ARISTAS COMO LINKS. LA PROBABILIDAD DE CADA ARISTA ES 1 SOBRE EL TOTAL DE LINKS EN LA PAGINA. PEDO INICIALMENTE TODAS SON EQUIPROBABLE-

EJEMPLO

- TENEMOS 5 PAGINAS
- INICIALMENTE TODOS LOS PESOS SON EQUIPROBABLES.



| A | B | C | D | E |
|-----|-----|-----|-----|-----|
| 1/5 | 1/5 | 1/5 | 1/5 | 1/5 |

AHORA REALIZO LA PRIMERA ITERACIÓN PARA CALCULAR LOS NUEVOS PESOS

$$A = \frac{1}{2} * \frac{1}{5} + \frac{1}{5} = \frac{3}{10} \rightarrow \text{LOS NUEVOS PESOS SE CALCULAN CON LO QUE RECIBE CADA NODO}$$

$$B = \frac{1}{2} * \frac{1}{5} + \frac{1}{5} = \frac{3}{10}$$

$$C = \frac{1}{2} * \frac{1}{5} + \frac{1}{2} * \frac{1}{5} = \frac{2}{10}$$

$$D = \frac{1}{2} * \frac{1}{5} = \frac{1}{10}$$

$$E = \frac{1}{2} * \frac{1}{5} = \frac{1}{10}$$

POR EJEMPLO A RECIBE DE B Y DE E - PERO B ENTRA EN DOS NODOS ≠ ENTONCES A RECIBE LA MITAD DE B Y TODO E

AHORA VOLVEMOS A HACER LO MISMO PERO CON LOS NUEVOS PESOS.

| A | B | C | D | E |
|------|------|------|------|------|
| 1/5 | 1/5 | 1/5 | 1/5 | 1/5 |
| 3/10 | 3/10 | 2/10 | 1/10 | 1/10 |

$$A = \frac{1}{2} * \frac{3}{10} + \frac{1}{10} = \frac{5}{20} \rightarrow \text{NOTAR QUE LAS ECUACIONES SON LAS MISMAS, SIMPLEMENTE CAMBIAN LOS PESOS.}$$

$$B = \frac{1}{2} * \frac{3}{10} + \frac{2}{10} = \frac{7}{20}$$

$$C = \frac{1}{2} * \frac{3}{10} + \frac{1}{2} * \frac{1}{10} = \frac{4}{20}$$

$$D = \frac{1}{2} * \frac{3}{10} = \frac{3}{20}$$

$$E = \frac{1}{2} * \frac{1}{10} = \frac{1}{20}$$

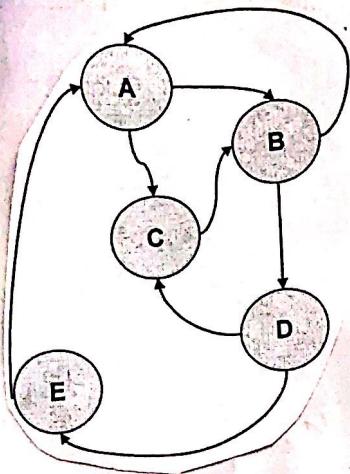
| A | B | C | D | E |
|------|------|------|------|------|
| 1/5 | 1/5 | 1/5 | 1/5 | 1/5 |
| 3/10 | 3/10 | 2/10 | 1/10 | 1/10 |
| 5/20 | 7/20 | 4/20 | 3/20 | 1/20 |

ASI REALIZAMOS VARIAS ITERACIONES HASTA QUE CONVERGA, OBTENIENDO EL PAGE RANK FINAL.

| | | | | |
|------|------|------|------|------|
| 6/25 | 8/25 | 5/25 | 4/25 | 2/25 |
|------|------|------|------|------|

ESTE MISMO SE PUEDE REPRESENTAR CON LA MATRIZ DE LA CADENA DE MARKOV

- LAS COLUMNAS SON LOS LINKS QUE SALEN DE CADA PAGINA.
- LAS FILAS SON LAS PAGINAS EN SI



$$M = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 1 \\ 1/2 & 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 \end{pmatrix}$$

NOTAR QUE ESTA COLUMNA SON LOS LINKS DE E. COMO VEMOS EN LA FIGURA, E SOLO TIENE LINK A A.

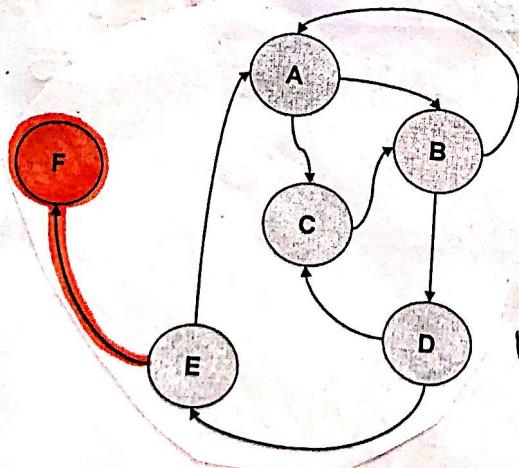
Y LOS PESOS COMO SABEMOS INICIALMENTE SON IGUALES LO TOMAMOS COMO UN VECTOR $(1/5 \ 1/5 \ 1/5 \ 1/5 \ 1/5)$. PARA CALCULAR EL NUEVO PESO HAY QUE MULTIPLICAR LA MATRIZ M POR EL VECTOR PESO.

$$\begin{pmatrix} 0 & 1/2 & 0 & 0 & 1 \\ 1/2 & 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \end{pmatrix} = \begin{pmatrix} 3/10 \\ 3/10 \\ 1/5 \\ 1/10 \\ 1/10 \end{pmatrix}$$

ESTOS SON LOS NUEVOS PESOS

AHORA SIMPLEMENTE MULTIPLICO M POR LOS NUEVOS PESOS Y ASÍ SUCESSIVAMENTE HASTA QUE CONVERGA.

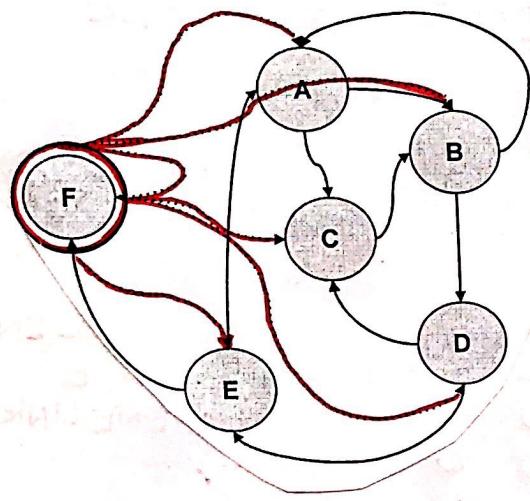
DEAD ENDS



LA PAGINA F ES UN DEAD-END, ES DECIR UNA PAGINA DESDE LA CUAL NO SE PUEDE IR A NINGUNA OTRA PAGINA. EN ESTA SITUACIÓN EL NAVEGANTE EVENTUALMENTE CAERÁ EN F Y QUEDARA ATRAPADO EN ESA PÁGINA

| A | B | C | D | E | F |
|--------|--------|--------|--------|--------|--------|
| 0,0018 | 0,0027 | 0,0017 | 0,0014 | 0,0007 | 0,9916 |

PARA CORREGIR ESTE ERROR, AL DEAD-END SE LE AGREGAN ENKS A TODAS LAS PAGINAS CON PROBABILIDAD $1/N$, INDICANDO AL NAVEGANTE QUE PUEDE VISITAR CUALQUIERA DE LAS PAGINAS.



ENTONCES DE TENER:

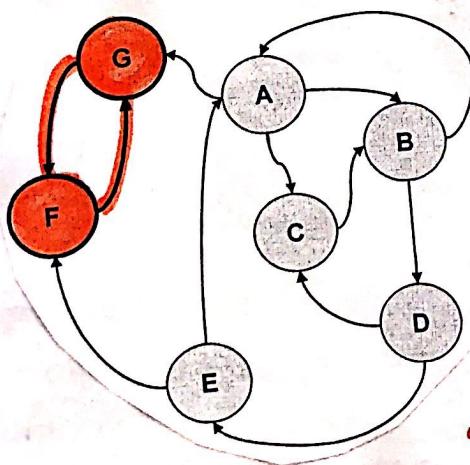
$$\left(\begin{array}{cccccc} 0 & 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right) = M$$

ESTO NO PUEDE SUCEDER, ES EL DEAD END.

= ESTO HAY QUE DESPLAZARLO POR LA COLUMNA TODA DE $1/6$ Y LUEGO OPERAR NORMALMENTE. LUEGO DE VARIAS OPERACIONES OBTENEMOS.

| A | B | C | D | E | F |
|--------|--------|--------|--------|--------|--------|
| 0,2044 | 0,3021 | 0,1912 | 0,1600 | 0,0889 | 0,0533 |

SPIDER TRAPS



PODEMOS OBSERVAR DEL GRAFO QUE UNA VEZ QUE NUESTRO NAVEGANTE LLEGUE A F VA A QUEDAR ATRAPADO EN F Y G SIN PODER SALIR. ESTO ES SPIDER-TRAP.

| A | B | C | D | E | F | G |
|-------|-------|-------|-------|-------|-------|-------|
| 0,014 | 0,018 | 0,011 | 0,010 | 0,006 | 0,461 | 0,481 |

LA SOLUCIÓN A ESTE PROBLEMA ES
TELETRANSPORTACIÓN

CADA VEZ QUE NUESTRO NAVEGANTE LLEGUE A UNA PAGINA, VAMOS A DARLE DOS POSIBILIDADES:

- CON PROBABILIDAD β VA A ELEGIR UN LINK AL AZAR. (COMO VENIAMOS HACIENDO)
 - CON PROBABILIDAD $1-\beta$ VA A TELETRANSPORTARSE A CUALQUIER OTRA PÁGINA (SIENDO EQUIPROBABLE)
- ENTONCES ANTES DE APLICAR LA TELETRANSPORTACIÓN NUESTRA MATRIZ ERA:

$$\left(\begin{array}{ccccccc} A & B & C & D & E & F & G \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right) = M$$

AHORA LA NUEVA MATRIZ M' APlicando la TRANSPORTACIÓN VA A SER DE LA FORMA.

$$M' = \beta M + (1-\beta) N$$

SIENDO N LA MATRIZ CON TODOS SUS VALORES $1/N$.

CON ESTA MATRIZ, HAGO EL MISMO PASO QUE ANTES PARA CALCULAR LOS PESOS.

→ A SU VEZ LA TELETRANSPORTACIÓN SOLUCIONA OTRO PROBLEMA, LOS GRAFOS PERIODICOS.

INTERPRETANDO β

β ES LA PROBABILIDAD DE SEGUIR UN LINK ES DECIR DE NO TELETRANSPORTARSE. POR LO GENERAL $0,8 < \beta < 0,9$.

- $\beta = 1$ ES EL CASO SIN TELETRANSPORTACIÓN
- $\beta < 1$ LA CADENA SE DEINICA CADA TANTO, CUANTO MENOR SEA, MAS RANDOM WALKS VAMOS A TENER Y MAS COORTOS VAN A SER.

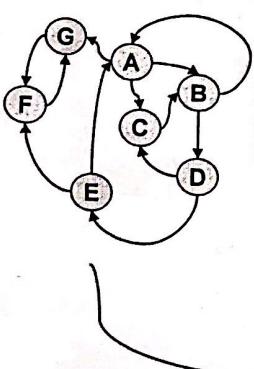
PAGE RANK - MAP REDUCE

SI TENEMOS UNA MATRIZ CON BILLONES DE NODOS NO PODEMOS REALIZAR LO QUE VIENIAMOS HACIENDO.

EXSISTE UNA RELACION SIMBIOTICA ENTRE PAGE RANK Y MAP REDUCE, ESTO EXPLICA EL GRAN EXITO DE GOOGLE, YA QUE M.R. SUBSE COMO SOLUCION AL PROBLEMA DE APLICAR PR A TODA LA WEB.

MAP

ESTE PROCESO, PROCESA UNA PAGINA CON SUS LINKS Y EMITE UN REGISTRO POR CADA LINK TRANSFIRIENDOLE LA PARTE DE PR QUE CORRESP.



CANTIDAD DE NODOS A LOS QUE APUNTA

| | | |
|---|-----|-----------|
| A | 1/7 | (B, C, G) |
| B | 1/7 | (A, D) |
| C | 1/7 | (B) |
| D | 1/7 | (C, E) |
| E | 1/7 | (A, F) |
| F | 1/7 | (G) |
| G | 1/7 | (F) |

$$B * (1/N)$$

C

| | |
|---|--------------|
| B | 0,85*(1/7)/3 |
| C | 0,85*(1/7)/3 |
| G | 0,85*(1/7)/3 |

| | |
|---|--------|
| B | 0,0405 |
| C | 0,0405 |
| G | 0,0405 |
| A | 0,0607 |
| D | 0,0607 |
| B | 0,1214 |
| C | 0,0607 |
| E | 0,0607 |
| A | 0,0607 |
| F | 0,0607 |
| G | 0,1214 |
| F | 0,1214 |

ESTO SE HACE PARA TODOS LOS NODOS.

| | |
|---|--------|
| B | 0,0405 |
| C | 0,0405 |
| G | 0,0405 |
| A | 0,0607 |
| D | 0,0607 |
| B | 0,1214 |
| C | 0,0607 |
| E | 0,0607 |
| A | 0,0607 |
| F | 0,0607 |
| G | 0,1214 |
| F | 0,1214 |

$$A \quad (0,0607, 0,0607) \rightarrow A \quad 0,1214 + 1/7 * 0,15 = 0,142857$$

| | |
|---|-------------|
| A | 0,142857143 |
| B | 0,183333333 |
| C | 0,122619048 |
| D | 0,082142857 |
| E | 0,082142857 |
| F | 0,203571429 |
| G | 0,183333333 |

= HAGO LA TELETRANS.

$$\Sigma \text{VALORES} + (1-\beta) \frac{1}{N}$$

REDUCE

EN ESTE PROCESO SE SUMAN TODOS LOS PUNTAJES OBTENIDOS Y SE LE AGREGA LA TELETRANSPOZACIÓN.

4

APLICACIONES

TOPIC RANK

EN EL ALGORITMO TRADICIONAL NOS TELETRANSMITIMOS CON IGUAL PROBABILIDAD A CUALQUIER PÁGINA, PERO EN ESTE CASO, NOS TELETRANSMITIMOS SOLO A AQUELLAS PÁGINAS QUE TRABAJAN SOBRE UN DETERMINADO TEMA.

SUPONIENDO QUE SABEMOS CUÁLES SON LAS PÁGINAS QUE HABLAN DE DETERMINADO TEMA ENTONCES, EN LA TELETRANS. EN LUGAR DE SUMAR $(1-B) \frac{1}{N}$ A CADA PÁGINA VAMOS A SUMAR:

MEJOR!

$$(1-B) * \frac{1}{C}$$



EN LAS POS
DE LOS ELEM
QUE QUEREMOS.

SIENDO C, LA
CANTIDAD DE PÁGINAS
QUE ESTAN CLASIFICADAS
DENTRO DEL TEMA QUE
NOS INTERESA.

- SI QUEREMOS DARLE MAYOR PESO A LAS PÁGINAS TEMÁTICAS TENEMOS QUE DARLE MAYOR PROBABILIDAD A LA TELETRANSPOZACIÓN, ES DECIR, USAR UN B MÁS CHICO.

TRUST RANK

EL TRUST RANK LOGRA DISMINUIR EL EFECTO DE LAS SPAM FARMS. EL OBJETIVO DE ESTAS ES AUMENTAR EL PR DE UNA DETERMINADA PÁGINA. PARA LOGRAR ESTO LOS SPAMMERS CREAN UNA GRAN CANTIDAD DE PÁGINAS QUE LINKEAN A LA PÁGINA EN CUESTIÓN Y LINKS DESDE LA PAG A ESTAS OTRAS FORMANDO UNA 'RED'. LUEGO INSERTAN LINKS EN PÁGS PÚBLICAS QUE LLEVEN A LA PAG OBJETIVO.

LA FORMA EN LA QUE TRUST BANK COMBATE EL SPAMFARM ES ÚNICAMENTE TELETRANSPORTANDONOS A PÁGS CONFIABLES. ESTO ES EQUIVALENTE A TENER UN TOPIC BANK PERO EN ESTE CASO EL TEMA ES LA PAG CONFIABLE.

¿QUE PÁGS SON CONFIABLES?

ESTAS SON POR SELECCIÓN MANUAL O AQUELLAS TERMINADAS EN .GOV, .EDU, .MIL... ETC...

$$\text{SPAM MASS} = \frac{\text{PR} - \text{TR}}{\text{PR}}$$

A PARTIR DE ESTO PODEMOS ESTABLECER CIERTO UMbral Y ELIMINAR LAS PÁGINAS QUE ESTEN POR ENCIMA DE ESO.

SIM RANK

EN ESTE CASO SOLO NOS TRANSPORTAMOS A UNA ÚNICA PÁG. LO QUE OBTENEMOS ES UNA SERIE DE RANDOM WALKS EN LOS CUALES SIEMPRE VOLVEMOS AL MISMO PUNTO DE PARTIDA. EL PR DE CADA PÁG REPRESENTA LA PROBABILIDAD DE LLEGAR A CADA PÁGINA DESDE UNA CIERTA PÁG ORIGEN.

VIA MONTECARLO

EL PROBLEMA DEL ALGORITMO SIMPLÍCITO ES QUE NO PODEMOS CALCULAR EL SR DE CADA PÁGINA, ESTO TOMARÍA UN TOTAL DE N EJECUCIONES LO CUAL NO ES ACEPTABLE. CUANDO QUEREMOS OBTENER LAS PÁGINAS MÁS SIMILARES A ALGUNA HACEMOS UNA SIMULACIÓN DE RANDOM WALKS COMENZANDO EN LA PÁGINA Y CON UN CERTO P DONDE $1-P$ ES LA PROBABILIDAD DE RE-INICIAR EL RANDOM WALK. A MEDIDA QUE REALIZAMOS LOS RANDOM WALKS REGISTRAMOS POR QUÉ PÁGINAS PASAMOS Y CUANTAS VECES VISITAMOS CADA UNA. LUEGO DE SIMULAR UNA BUENA CANTIDAD DE RANDOM WALKS, Y ASÍ, TENEMOS UNA MUY BUENA APROXIMACIÓN AL SR DE LA PÁG SIMPLEMENTE LISTANDO LAS DEMAS EN ORDEN DECREciente POR CANTIDAD DE VISITAS.

ESTE METODO SE PUEDE USAR EN REDES SOCIALES PARA RECOMENDAR AMIGOS O USUARIOS A SEGUIR.

VISUAL RANK

VISUAL BANK ES UN ALGORITMO PARA BANKEAR IMAGENES EN UN BUSCADOR DE IMAGENES. ENTRE LAS IMAGENES QUE TIENEN EN SU TITULO LA FRASE BUSCADA PODEMOS TENER LAS MAS RELEVANTES. Y LAS MENOS RELEVANTES AL MISMO TIEMPO.

VISUAL BANK DESUELVE ESTE PROBLEMA CON LA IDEA: LA IMAGEN MAS RELEVANTE DEBE SER AQUELLA QUE MAS SE PARECE A TODAS LAS DEMAS. SI PODEMOS BANKEAR LAS IMAGENES, DE ACUERDO A SU SEMEJANZA CONTDA TODO EL RESTO TENDREMOS UN ORDEN MUY EFECTIVO PARA PRESENTARLE LOS RESULTADOS AL USUARIO.

SE COMIENZA ADMENDO UN GRAFO ENTRE TODAS LAS IMAGENES RECUPERADAS, DONDE CADA ARISTA REPRESENTA LA SEMEJANZA ENTRE LAS IMAGENES. USANDO LA SEMEJANZA COMO LA PROBABILIDAD DE VISITAR UNA LUEGO DE OTRA, PODEMOS BANKEAR LAS IMAGENES.

SIFT ES EL ALGORITMO QUE SE UTILIZA PARA CALCULAR LA SEMEJANZA ENTRE DOS IMAGENES. ESTE EXTRADE DE CADA IMAGEN UN CONJUNTO DE PUNTOS QUE NO SE MODIFICAN CON DIFERENTES ESCALAS, ROTACIONES O ILUMINACIÓN.

LA SEMEJANZA ENTRE DOS IMAGENES SE CALCULA MEDIANTE LA DISTANCIA ENTRE TODOS LOS PUNTOS SIFT DE UNA IMAGEN CONTDA TODOS LOS PUNTOS DE OTRA. COMO ESTO ES MUY COSTOSO, SE UTILIZA LSH PARA PODER REALIZAR LAS COMPARACIONES EN $O(1)$. LA SEMEJANZA ES SIMPLEMENTE LA CANTIDAD DE COLISIONES QUE TUVIMOS EN TOTAL DIVIDIENDO POR EL TOTAL DE PUNTOS SIFT.

TEXT RANK

ESTE ALGORITMO REALIZA DESUMENES AUTOMATICOS
UN DESUMEN AUTOMATICO CONSISTE EN EXTRAER DE UN
TEXTO SUS FRASES MAS SIGNIFICATIVAS. PARA ESTO
TEXT RANK ASEA UN GRAFO CON TODAS LAS FRASES DEL
TEXTO Y USA ALGUNA METRICA DE DISTANCIA PARA CALCULARLA
CON EL GRAFO CONSTRUIDO, TEXT RANK CORRE PR SOBRE
EL MISMO Y LUEGO EXTRAER LAS K FRASES DE MAYOR
PR PARA REALIZAR EL DESUMEN.